ICC 2017
WASHINGTON
JULY 02-07, 2017

Supervised Machine Learning for Regionalization of Environmental Data: Distribution of Uranium in Groundwater in Ukraine

Michael Govorov[1], Gennady Gienko[2], Viktor Putrenko[3]

1. Vancouver Island University, Nanaimo, BC, Canada; govorovm@viu.ca
2. University of Alaska Anchorage, Anchorage, AL, USA; ggienko@alaska.edu
3. National Technical University of Ukraine, Kiev, Ukraine; putrenko10@gmail.com

Abstract*:* In this paper, several supervised machine learning algorithms were explored to define homogeneous regions of concentration of uranium in surface waters in Ukraine using multiple environmental parameters. The previous study was focused on finding the primary environmental parameters related to uranium in ground waters using several methods of spatial statistics and unsupervised classification. At this step, we refined the regionalization using Artificial Neural Networks (ANN) techniques including Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Convolutional Neural Network (CNN). The study is focused on building local ANN models which may significantly improve the prediction results of machine learning algorithms by taking into considerations non-stationarity and autocorrelation in spatial data.

Keywords: Deep Machine Learning, Regionalization, Spatial Effect, Uranium, Groundwater

1. Introduction

One of the main components of ecological conditions in Ukraine is the radioactivity of natural waters associated with natural and anthropogenic factors. Surface and ground waters are important resources of drinking water. Several areas in Ukraine have high concentrations of natural uranium, so surface and groundwater in these areas can be potentially unsafe as a source of drinking water. Concentrations of uranium of 0.08 Mg/L and higher are potentially dangerous to human health, therefore, investigation of the impacts of uranium on groundwater (and thus, on the quality of drinking water) is an important scientific problem.

The goal of this study is to explore the application of several supervised machine learning algorithms to define homogeneous regions of concentration of uranium in surface waters in Ukraine using labeled training data and multiple environmental variables. In this study, regionalization is defined as a generalization of properties of a phenomenon throughout space, based on a set of training examples and a set of various observations. The regionalization of Ukraine by the concentration of uranium in groundwater allows defining quality standards for drinking water for the acceptable content of uranium and associated elements. Also, the regionalization can be used as a new approach to conducting geological work to search for mineral deposits.

2. Methodology

To enhance and validate the proposed in previous studies regionalization models, the authors explore several methods of supervised learning, including Artificial Neural Networks techniques such as Multilayer Perceptron

(MLP) and Radial Basis Function (RBF) networks as functions of environmental parameters which minimize the prediction error of dependent variable defined group membership. Convolutional Neural Network (CNN) which is a relatively new enhanced version of MLP, has also been explored (Haykin 2011, Goodfellow et al. 2016).

A Multilayer Perceptron (MLP) is the type of ANN that contains one or more hidden layers of neurons or nodes (apart from one input and one output layer). MLP learns or is trained by using the backpropagation algorithm that is a supervised training scheme from labeled training data. Conventional MLP can be applied to perform regionalization of spatial data without modification. However, spatial data require some treatment of spatial autocorrelation and nonstationarity.

There are at least three spatially-based clustering approaches which take into account the spatial effect (Simbahan and Dobermann 2006, Hu and Sung 2004). The first approach is to add spatial information into datasets. The simplest way is to use geographic coordinates as additional classification variables and achieve spatial contiguity by assigning an appropriate weight to the geographic coordinates (Webster and Burrough, 1972, Govorov, 1986). In spatially weighted classification, the principal coordinates of dissimilarity matrices are modified spatially, which then can be used to create classifications. The dissimilarity measure can be weighted as a function of the geographic separation between individuals to ensure spatial continuity of the formed clusters, for example, incorporation of known autocorrelation among data from uni- or multivariate variograms into their spatial classification (Bourgault et al., 1992). Principal component semivariograms can be used instead of variable-specific semivariograms.

The second approach is modifying existing algorithms, e.g., contiguity-constrained classification. Spatial contiguity can be modeled by creating a set (or cluster) of neighbors for each sample point. Contiguity-constrained classification imposes a constraint, which determines individuals or groups that can be joined to form a cluster (Openshaw, 1977, Ferligoj and Batagelj, 1982, Davidson and Basu 2007, Duque 2007). The third approach is to build a model that encompasses spatial information.

In terms of MLP models, the first and second approaches can be used to incorporate spatial information into the network propagation. Thus, weighted coordinate values can be used as independent variables, as well as input independent variables can be spatially weighted, e.g., each input can be recalculated based on the weighted average of its neighbors and used as new inputs to a conventional MLP network. As an alternative to input filtering on independent variables, conventional MLP network can be trained, and then predicted values from the MLT output layer can be modified based on a spatial filter, similar to post-classification smoothing (Simbahan and Dobermann, 2006).

One more approach to consider spatial effect in the MLP model is to incorporate spatial autocorrelation into the hidden layer(s) by modifying input associated weights that are used to output from hidden layer to output layer. Each input has an associated weight, which is assigned on the basis of its relative importance to other inputs. The goal of learning is to assign correct weights for the connections between neurons of adjacent layers. Conventionally all the neuron connection weights are randomly assigned. To train a network, the ANN is activated for every input in the training data set and calculates its output by using an activation function. An activation function can be modified to consider autocorrelation between the neighboring neurons. This output is compared with the known labeled output, and the error is propagated back to the previous layer. The weights are adjusted accordingly to this error by using an optimization method. This process is repeated until the output error is below a predetermined threshold. Then the trained network can be used to classify new inputs.

Another ANN model, which can be used for spatial classification, is radial basis function networks (RBF networks) (Yee and Haykin 2001, Que and Belkin 2016). A special class of radial function can be employed as activation function in a single or multi-layer network. Traditionally, RBF is used in a single layer network where the optimal subset of radial basis functions is used for activation in forward selection. Activation process fits radial basis functions with weight to the hidden layer's outputs with respect to some objective function. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. A backpropagation step can be performed to adjust the RBF network parameters.

Conventional RBF networks assume the independent and identical distribution of input variables. To work with spatial data, RBF can be adjusted by using similar approaches as for MLP model: by filtering data values of input or output layer; by incorporating spatial autocorrelation into the output weight from hidden radial basis layer by modi-

fying the linear combination of RB functions. RB functions can be modified similarly to spatial econometrics autoregressive models to model substantive spatial dependence (Anselin 1995).

MLP and RBF are fully-connected networks where each neuron is connected to every neuron in the previous layer, and each connection has its own weight. This is a general purpose connection pattern and makes no assumptions about the autocorrelation in the spatial data. For cases where the data can be interpreted as spatially correlated, Convolution Neural Networks (CNN) can be employed. CNN intends to use spatial information between the neighboring neurons are based on discrete convolution (Zeiler and Fergus 2013).

CNN may have any number of convolution, normalization and pooling layers after the input layer. The output from the last pooling layer acts as an input to the fully connected layer of CNN. The convolution/normalization/pooling layers act as a filter(s) for object extraction from the input data while fully connected layer acts as a classifier. Formally, CNN is similar to MLP or RBF where the input data are spatially filtered, or thus fully connected layer is a convolutional layer with filter size equal to input size.

However, in a convolutional layer, each neuron is only connected to a few neighboring neurons in the previous layer, and weights are assigned only to this local connections. Convolution preserves the spatial relationship between neurons. During the training process, CNN learns its weights and filter values and then adjusting them during backpropagation process. However, CNN parameters such as the number of filters, filter sizes, the architecture of the network etc. have all been fixed and do not change during the training process. Filters with different sizes can be used for different spatial locations.

CNN can be applied not only on regular raster data but also irregular point data where natural neighbors can be found e.g. based on Delaunay triangulation and used on the convolution steps to extract features from the input data.

3. Case Study

In the previous studies (Govorov et. al 2016), the authors used several spatial statistical methods including exploratory spatial data analysis, global and local factor analysis (proposed geographically weighted factor analysis), correlation and regression analysis (geographically weighted correlation and regression analysis), autoregressive models of spatial econometrics were utilized to describe the impact of several environmental variables on spatial distribution of uranium (U) in ground water. As the result, it was found that concentration of uranium has a strong local correlation with precipitation, humus, the hardness of water, F, Fe, SO_4 and As. These first six most significant predictors contribute 60.79% into the overall regression model (Table 1). Since some of these source variables are dependent the principal components of these variables can be used for analysis.

Table 1. Multiple regression U vs. 17 variables model summary

Model	Predictors	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	precipitation	0.510	0.260	0.260	1.1611
2	1 + humus	0.595	0.354	0.354	1.0849
3	2 + water hardness	0.623	0.388	0.387	1.0564
4	3 + F	0.629	0.396	0.395	1.0496
5	4 + Fe	0.634	0.402	0.402	1.0438
6	5 + As	0.637	0.405	0.405	1.0414
7	6 + SO_4	0.638	0.407	0.407	1.0395
8	7 + isopach	0.641	0.411	0.410	1.0368
9	8 + NH_4	0.642	0.413	0.412	1.0351
10	9 + Cl	0.644	0.415	0.414	1.0332

11	10 + temperature	0.645	0.416	0.415	1.0325
12	11 + NO_3	0.645	0.417	0.416	1.0317
13	12 + HCO_3	0.646	0.418	0.416	1.0310
14	13 + Zn	0.647	0.418	0.417	1.0305
15	14 + Cu	0.647	0.419	0.418	1.0300
16	15 + PO_4	0.648	0.419	0.418	1.0296
17	16 + mineralization	0.648	0.420	0.419	1.0288

Then, several unsupervised machine learning algorithms were explored to define homogeneous regions of concentration of uranium in ground waters using multiple environmental parameters. At this step, cluster analysis was carried out using techniques of bivariate local pattern analysis, spatially weighted classification and spatially contiguous clustering of multivariate data or unsupervised learning, and techniques from the domain of artificial neural network, specifically Kohonen Self-Organizing Maps was used (Kohonen, 2001).

Combining techniques of hierarchical and non-hierarchical classification of geological, climatic, and various environmental parameters, coupled with geostatistical analysis, allowed for the creation of several regionalization maps for the study area. The maps from different classification and clustering algorithms have been analyzed based on several criteria, including spatial homogeneity.

Analysis of geographic and climatic features, along with the geological structure of the area, prompted to implement a hierarchical approach for zoning, which resulted in identifying three main agglomerated areas with different conditions of accumulation of uranium in groundwater (Figure 1, left) (Makarenko 2000, Kirovgeologija 2004). The main approach was to reflect the association of climatic variables with the tendency of increased mineralized groundwater towards the southeast of the country. The three outlined areas were further subdivided into six zones (Figure 1, right) based on detailed analysis of different models. The resulting zoning has a hierarchical structure which makes it more flexible and evolutionary adaptive in making decisions for geological studies, environmental assessment, and the use of groundwater.

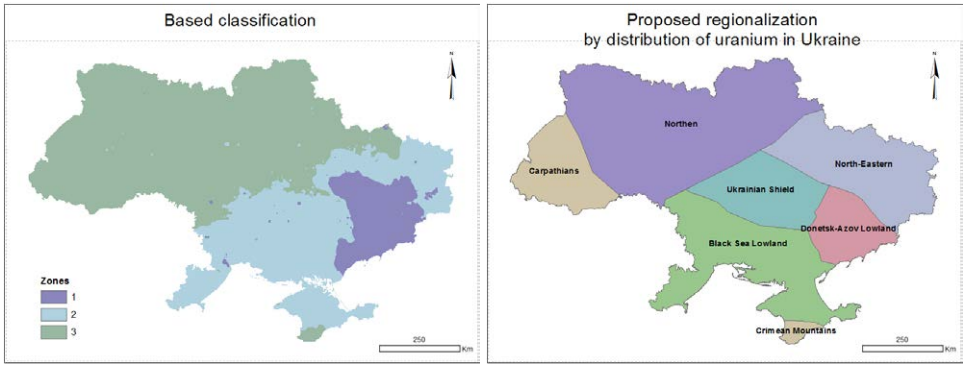

Fig. 1. Three main zones with different conditions of accumulation of uranium in groundwater (left) and proposed regionalization of distribution of uranium in Ukraine (right)

However, the proposed spatial classification (regionalization) was based on more than 20 clusterization outputs/maps, which show cognitive clusters, but at the same time, these maps were rather different. The applied validation techniques did not provide an unbiased answer to the question: what the most reliable clustering output is.

Applying ANN supervised learning techniques to the unsupervised classifications from the previous experiments resulted in substantially enhanced regionalization. Validation and comparison of classification results were performed by using three approaches. First, external indices have been used to measure the extent to which cluster labels match externally supplied class labels. Second, internal indices were used to measure the goodness of a cluster-

ing structure without respect to external information (overall similarity), and the third approach is based on the use of a relative index to compare two different clustering results.

The analysis was implemented using SPSS with R extensions, ArcGIS, and MatLab. The study resulted in a series of refined maps which show homogeneous regions of primary environmental variables (or their principle components) based on their relationships with the concentration of natural uranium in ground waters. Regionalization maps, which were obtained by using global and local modifications of MLP, RBF and CNN models, will be presented on the respective oral session during the ICC2017 conference.

3. Conclusions

In this studies, the authors investigated several ANN techniques for classification of environmental spatial data represented as a point cloud. There are a few ways to adopt ANN to work with spatial data, including CNN which has been already designed to use spatial information between the neighboring neurons based on discrete convolution. However, in order to use CNN for irregular point classification, a regular matrix of neighbors can be constructed, and thus the appropriate architecture of the network, and parameters for convolution (e.g., type, size, and a number of filters, etc.) and pooling (subsampling) layers should be designed. For example, filter size can be selected as peaks that reflect distances where the spatial processes promoting most pronounced clustering which can be calculated based on Global Moran's autocorrelation index (Moran 1950, Ord and Getis 2001). The further study is aimed at exploring several approaches which incorporate spatial constraints into networks. For example, semi-supervised clustering that is combining ANN techniques (e.g., RBF) and clustering algorithms (e.g., spatially constrained K-means) has a potential to increase the accuracy of regionalization.

References

Anselin, L (1995) Local Indicators of Spatial Association – LISA. Geographical Analysis, 27(2), 93–115.

Bourgault, G., Marcotte, D., Legendre, P. (1992) The multivariate (co) variogram as a spatial weighting function in classification methods. Math. Geol. 24, 463–478.

Davidson, I, Basu, S (2007) A Survey of Clustering with Instance Level Constraints. ACM Transactions on Knowledge Discovery from Data.

Duque, J.C., Ramos, R., Surinach, J. (2007) Supervised Regionalization Methods: A Survey, International Regional Science Review 30: 195–220

Ferligoj, A., Batagelj, V. (1982) Clustering with relational constraint. Psychometrika. 47, 413–426.

Getis, A., Ord, J.K. (1992) The analysis of spatial association by use of distance statistics, Geographical Analysis. 24(3), 186-206

Goodfellow, I., Bengio, Y., Courville, A. (2016) Deep Learning, MIT Press.

Govorov M, Putrenko V, Gienko G (2016) Mining Spatial Patterns of Distribution of Uranium in Surface and Ground Waters in Ukraine, Handbook of Research on Geographic Information Systems Applications and Advancements, Chapter 21, IGI Global.

Govorov, M.O., Malikov, B.N. (1986) Selected tropics of the use of geological and economic maps. Geography and Natural Resources, 4, 165-167.

Haykin, S.O. (2011) Neural Networks and Learning Machines, 3rd Edition, Kindle Edition.

Hu, T., Sung, S.Y. (2004) Predicting spatial data with RBF networks. International Journal Neural Systems, 14(2):117-23.

Kohonen, T. (2001) Self-Organizing Maps, 3rd ed. Springer, Berlin.

Kirovgeologija (2004) Prirodni ta antropogenni dzherela formuvannja radioaktivnosti prirodnih vod Ukraïni ta radiacijnij zahist naselennja. (2004). Derzhkomprirodresursiv Ukraïni, Departament geologichnoï sluzhbi, Kazenne pidpriємstvo "Kirovgeologija", Kiev.

Makarenko, M.M. (2000) Ocinka prirodnih i tehnogennih faktoriv zabrudnen' pidzemnih i poverhnevih vod prirodnimi radionuklidami navkolo uranovih rodovishh Ukraïni. Informacijnij bjuleten' pro stan geologichnogo seredovishha Ukraïni (pp. 102-111), Kiïv

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. Biometrika, 37, 17-23.

Openshaw, S. (1977) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. Trans. Inst. Br. Geogr. 2, 459–472.

Ord, J.K., Getis, A. (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. J Reg Sci 41(3):411-432.

Que, Q., and Belkin, M. (2016) Back to the future: Radial Basis Function networks revisited. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS).

Simbahan, G.C., Dobermann, A. (2006) An algorithm for spatially constrained classification of categorical and continuous soil properties, Geoderma, Volume 136, Issues 3–4, 15 December 2006, 504–523.

Webster, R., Burrough, P.A. (1972) Computer-based soil mapping of small areas from sample data: II. Classification smoothing. J. Soil Sci. 23, 222–234

Yee, P.V. and Haykin, S. (2001). Regularized Radial Basis Function Networks: Theory and Applications. John Wiley.

Zeiler, M.D., and Fergus, R. (2013) Visualizing and understanding convolutional networks. Computing Research Repository, abs/1311.2901.